A Text-to-SQL Benchmark for Official Statistics

Lucas Lageweg [1,2], Jan-Christoph Kalo [1], Karen Goes [2], Tom van den Broek [2], Frank Pijpers [2] and Paul Groth [1]

[1] INDElab, University of Amsterdam, The Netherlands
{l.lageweg,j.c.kalo,p.t.groth}@uva.nl
[2] Centraal Bureau voor de Statistiek, The Netherlands
{l.lageweg,kwm.goes,aj.vandenbroek,f.pijpers}@cbs.nl

Official statistics are an immensely valuable dataset resource, on which policy makers and businesses rely. Making official statistics easily accessible is one of the key tasks for the agencies making these resources available, for example by providing natural language question answering systems (i.e. text-to-SQL) [1]. Current text-to-SQL datasets, however, do not represent the complexity of the data found in official statistics. Popular 'large-scale datasets' such as Spider [4] and BIRD [2], are comparatively limited in size and often do not contain real-world enterprise data or only focus on a specific domain, limiting the benchmark in its scope. In official statistics, the task of QA concerns very high volumes of complex data ranging over multiple domains. Tables typically contain a lot of different measured concepts (observations), where these concepts are described using jargon and legal terms not often seen by LLMs or queried by end-users, requiring external knowledge from models to be able to interpret and answer correctly. Furthermore, when querying and aggregating values from tables in statistics, the calculation methods and units of these measures are essential in providing factual information.

We present LoTuS: a Large-scale Text-to-SQL benchmark for official Statistics. Our to-be-released benchmark contains over 2,200 Dutch and English tables containing real-world statistics, table schemas ranging from small (< 5 columns) to very large ($> 1,200$ columns), a coverage of 22 statistical domains, 2500 manually annotated complex questions and answers and an extensive metadata knowledge graph accompanying the tables, containing labels, descriptions, units, and other associated metadata for concepts. Preliminary results show an execution accuracy of 0.59 using a finetuned ColBERT [3] retriever and a zero-shot prompting strategy on OpenAI's GPT4.1, with and observation F1 of 0.64, indicating ample room for further research improvements. This benchmark will be a valuable resource for testing models in non-relational large-scale and multi-domain enterprise datasets, as well as providing a challenging task for Dutch language models alongside its more resource-abundant English counterparts.

Dataset	**# queries**	**# tables**	**# avg. cols/table**
Spider [4]	10,181	1,932	5,44
BIRD [2]	12,751	612	7,14
LoTuS-nl	1,250	1,211	34.55
LoTuS-en	1,250	1,030	32,40

Table 1: Comparison of dataset statistics for Spider and BIRD to LoTuS.

References

[1] European Parliament and Council of the European Union. 2009. Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics. Official Journal of the European Union, L 087, p. 164. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02009R0223-20241226

[2] Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Rongyu Cao, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C. C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023. Can LLM Already Serve as A Database Interface? A BIg Bench for Large-Scale Database Grounded Text-to-SQLs. https://doi.org/10.48550/arXiv.2305.03111 arXiv:2305.03111 [cs].

[3] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. arXiv:2112.01488 [cs.IR] https://arxiv.org/abs/2112.01488

[4] Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2019. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. https://doi.org/10.48550/arXiv.1809.08887 arXiv:1809.08887 [cs].